1. 学习雪花算法。
2. 学习递归算法
3. 成都分公司知文数据获取—贵州省科技厅

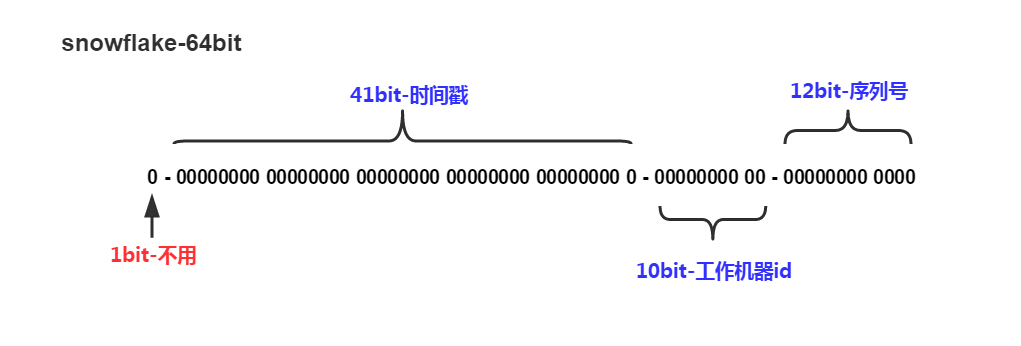
# 雪花算法

## 背景

雪花算法又名Twitter-Snowflake，该算法产生的背景比较简单，主要是为了满足Twitter每秒上万条消息的请求，而每条消息都必须分配一个唯一的id并且这些id还具体某种顺序（方便客户端排序），并且在分布式系统中不同机器产生的id必须不同。该算法属于划分命名空间并行生成的一种算法，生成的数据为64bit的long型数据，在数据库中应该用大于等于64bit的数字类型的字段来保存该值，比如在MySQL中应该使用BIGINT。雪花算法主要由四部分组成，但是真正起作用的只有后面三部分：时间戳，工作机器id，序列号。通常情况下是将这三部分组合在一起生成一个不唯一id。

## 算法原理

SnowFlake算法产生的ID是一个64位的整型，结构如下（每一部分用“-”符号分隔）：



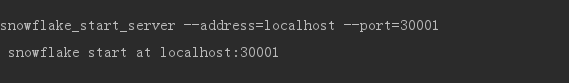
* 最高位是符号位，始终为0。
* 41位的时间序列：能够精确到毫秒级，41位的长度可以使用69年。该部分还有一个很重要的作用是可以根据时间进行排序。
* 10位的机器标识：Twitter实现中使用前5位作为数据中心标识，后5位作为机器标识，可以部署1024个节点
* 12位的计数序列号：序列号即一系列的自增id，可以支持同一节点同一毫秒生成多个ID序号，12位的计数序列号支持每个节点每毫秒产生4096个ID序号。

该算法通过时间戳和标识来防止重复，通过序列号实现自增。但是，不是所有的发号器都需要64位，可以根据项目的实际情况分配不同位数。如：若不需要数据中心，完全可以使用全部10位作为机器标识；若数据中心不多，也可以只使用3位作为数据中心，7位作为机器标识。、

pysnowflake的使用:

安装pysnowflak：pip install pysnowflak

启动pysnowflak服务：在命令行输入（除了指定ip和端口外，还可以指定数据中心和机器标志）



例如：

# --address：本机的IP地址默认localhost

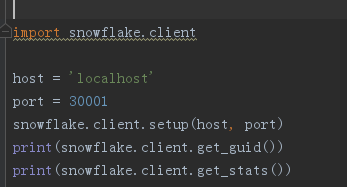
# --dc：数据中心唯一标识符默认为0

# --worker：工作者唯一标识符默认为0

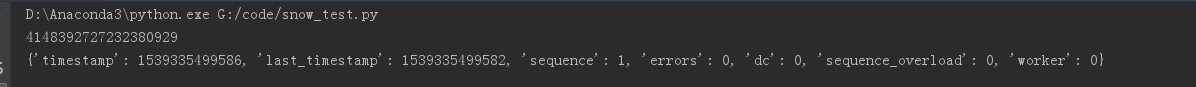
# --log\_file\_prefix：日志文件所在位置



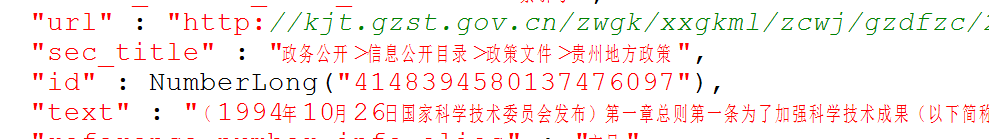
连接pysnowflake并获取id：



结果如下：



例如将该id存在mongodb中如下所示，该id是long型数据。



# 递归算法

## 定义

递归就是子程序（或函数）直接调用自己或通过一系列调用语句间接调用自己，是一种描述问题和解决问题的基本方法。

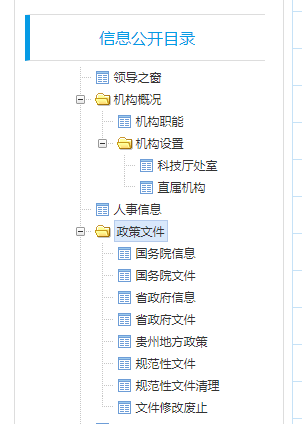
递归常与分治思想同时使用，能产生许多高校的算法。递归常用来解决结构相似的问题。所谓结构相似，是指构成原问题的子问题与原问题在结构上相似，可以用类似的方法解决。即就是整个问题的解决，可以分为两部分：第一部分是一些特殊情况，有直接的解法；第二部分与原问题相似，但比原问题的规模小，并且依赖第一部分的结果。实际上，递归是把一个不能或不好解决的大问题转化成一个或几个小问题，再把这些小问题进一步分解成更小的小问题，直至每个小问题都可以直接解决。因此，递归有两个基本要素：

（1） 边界条件：确定递归到何时终止，也称为递归出口。

（2） 递归模式：大问题是如何分解为小问题的，也称为递归体。

递归函数只有具备了这两个要素，才能在有限次计算后得出结果。

为什么会学习递归算法呢？主要是在爬取贵州省科技厅信息公开目录下的数据的时候，发现有很多子目录，如果利用轮询的方法程序就会很繁琐。



递归获取信息公开目录下符合要求模块的url代码如下：

